DOCUMENT RESUME

ED 052 232 TM 000 636

AUTHOR Lipe, Gary
TITLE Analysis of Performance Data for Instructional Design Projects.
SPONS AGENCY Office of Naval Research, Washington, D.C. Psychological Sciences Div.
PUB DATE Feb 71
NOTE 21p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York, New York, February 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS Autoinstructional Aids, Criterion Referenced Tests, *Data Analysis, Evaluation Techniques, *Instructional Design, Instructional Materials, *Learning, Literature Reviews, *Models, *Performance, Post Testing, Pretesting, Systems Approach, Test Construction, Testing

ABSTRACT

The development of a single value revision indicator which would utilize learner performance data obtained from a pretest-posttest design to rank a set of instructional modules as to their relative need for revision is discussed. A set of procedures was developed in connection with the implementation of the Production, Implementation, Evaluation, and Revision of Instruction Modules (PIERIM) Model for design of instruction. A comparison of the similarities and differences between using the module in a conventional classroom environment and using it in a self-instruction environment are presented as a frame of reference for the analysis and interpretation of the learner performance data reported in Tables 1 and 2. The correlation coefficient (r =.83) indicated substantial agreement between the rankings of the instructional modules using the revision indicators derived from the learner performance data for Group I and Group II. This methodology appears to be one method for the better utilization of data derived from learner performance during the formative evaluation of instructional materials. (CK)

Analysis of Performance Data for Instructional Design Projects[1]

by

Gary Lipe, Assistant Professor
School of Education
Texas Christian University

A Paper Presented to the
National Council on Measurement
in Education
February 4-7, 1971
New York, New York

1 This paper is based on research reported in Lipe, J.G. The Development and Implementation of a Model for the Design of Individualized Instruction at the University Level. CAI Center Tech Report No. 15, Florida State University, Tallahassee, Florida, 1970. The research was supported by Project NR 154-280, Sponsored by Personnel & Training Research Programs, Psychological Sciences Division, Office of Naval Research, Washington, D.C., Contract No. N00014-68-A-0494.

ANALYSIS OF PERFORMANCE DATA FOR INSTRUCTIONAL DESIGN PROJECTS

by Gary Lipe

1. PROBLEM

A selected review of the research related to the areas of criterion-referenced measures and evaluation provide a background for the discussion of the development of a Revision Indicator to be used in connection with the formative evaluation of instructional modules.

Criterion-Referenced Measures

Glaser (1963; 1967), Glaser and Cox (1968), and Popham and Husek (1969) discussed not only the similarities and differences between norm-referenced measures and criterion-referenced measures, but also the application of criterion-referenced measures to evaluation of instruction. A criterion-referenced test was operationally defined to include any measure which:

1. Assesses learner performance in relation to a predetermined standard of performance.
2. Provides information as to the level of performance by each learner which is independent of reference to the performance of other learners (after Glaser, 1968 and Glaser and Cox, 1968).

Ebel (1962) discussed ten principles which should be considered when tests of educational achievement were being prepared and used. The first five principles were considered to be equally applicable to criterion-referenced measures:

1. The measurement of educational achievement is essential to effective education.
2. An educational test is no more or less than a device for facilitating, extending, and refining a teacher's observation of student achievement.
3. Every important outcome of education can be measured.
4. The most important educational achievement is command of useful knowledge.
5. Written tests are well suited to measure the student's command of useful knowledge (p. 20-22).

Evaluation

Definition

Merwin (1969) reviewed the historical development and changing concept of evaluation and concluded that "concepts of evaluation have developed in response to needs for evaluational practices . . . (p. 25)." The combination of ideas from Stake's (1967) discussion of curriculum evaluation, Scriven's (1967) discussion of formative evaluation and Wittrock's (1969) discussion of evaluation of instruction resulted in the following definition:

> Formative evaluation is the collection, processing, and interpretation of data for the purpose of describing and making judgement as to the quality and appropriateness of behavioral objectives, instructional materials, environments, and learner performance, and utilizing the results to make decisions concerning the modification of the instructional system from which the data was derived.

Modification of a system based on data derived from the system (e.g., output) implies feedback. Feedback has generally been defined as any output of a system which either directly or indirectly serves as future input to the system. Within the context of a system model for design of instruction, the role of the evaluator is to utilize the output of the system to identify possible weaknesses within the system which, if corrected, would increase the efficiency of the total system and/or proportion of learners attaining the specified standard of performance. Feedback to the instructor provides the information required to make decisions concerning the modification of instructional materials and/or procedures (Bloom, 1968 , 1969; Cronback, 1963; Glaser, 1965; Tyler, 1949, 1951; Wittrock, 1969). The information can also be used to modify the product of any of the steps in a system model for design of instruction (Briggs, 1970; Dick, 1969).

There are few specific guidelines concerning the data to be collected, techniques for analyzing the data, or decision strategies for assigning priorities to the changes which must be made to an instructional system. Recommendations are reviewed for test items and instructional materials.

Test Items

System models for design of instruction and mastery models each identify the first concern in evaluation test items, which is to establish the content validity of the item (Bloom, 1968, Cronbach, 1963; Ebel, 1956; Husek, 1969; Popham & Husek, 1969; Tyler, 1949; Wittrock, 1969). When test items are derived directly from statements of behavioral objectives, as they are in a system model for design of instruction, the content validity of the item has been established.

Empirical testing of test items, using both individual and small group procedures, has been recommended by Tyler (1949). The method of scoring the performance of a learner should be made as objective as possible (Bloom, 1969; Lindvall & Cox, 1969; Tyler, 1949; Wittrock, 1969), and the basis of scoring should be made known to the learner (Wittrock, 1969). Evans (1968) recommended the use of multiple-choice type items whenever possible and contended that the ultimate operational definition of the instructional system's objectives is the posttest used to evaluate the learner's performance.

Cox and Vargas (1966), Glaser and Cox (1968), Hills (1970), Husek (1969), Moxley (1970), Popham (1970), and Popham and Husek (1969) have all expressed concern because of the lack of appropriate methods of analyzing data from criterion-referenced measures of learner performance. The suggested recommendations have been very general in nature, such as: the proportion of learners passing an item should be low on the pretest and high on the posttest (Glaser & Cox, 1968; Moxley, 1970), and a negative discriminator in an item pool should be carefully analyzed (Popham & Husek, 1969). Specific procedures for item analysis, based on the pretest-posttest design, have been discussed by Cox and Vargas (1966) (e.g., pretest-posttest difference index) and Popham (1970) (e.g., fourfold analysis of pretest-posttest learning states).

Instructional Material

The pretest-posttest design has been widely recommended and is essential if learning is to be inferred from changes in the learner's performance before and after interacting with an instructional system (Deterline, 1967; Glaser & Cox, 1968; Lindvall & Cox, 1969; Lumsdaine, 1965; Provus, 1969; Tyler, 1949; Wittrock, 1969). The pretest-posttest design is considered a minimal design by Tyler (1949) and additional observations of the learner's performance were recommended to estimate the retention of the performance. When the only data available to an evaluator is from a pretest-posttest design, it is exceedingly difficult to determine which element of the instructional system should be revised.

The problem was to develop a single value Revision Indicator which would utlize learner performance data obtained from a pretest-posttest design to rank a set of instructional modules as to their relative need for revision.

2. PROCEDURES

The following set of procedures were developed in connection with the implementation of the Production, Implementation, Evaluation, and Revision of Instructional Modules (PIERIM) model for design of instruction (Lipe, 1970). A comparison of the similarities and differences which existed during Phase 2--Implementation and Evaluation of Instructional Module in a Conventional Classroom Environment and Phase 4--Implementation and Evaluation of Instructional Modules in a Self-Instruction Environment of the PIERIM model provides a frame of reference for the analysis and interpretation of the learner performance data reported in Tables 1 and 2.

INSTRUCTIONAL MODULES	NUMBER ITEMS	MEAN		STANDARD DEVIATION	
		PRETEST	POSTTEST	PRETEST	POSTTEST
Pretest/Posttest	3	1.42	1.74	.59	.91
Behavioral Objectives	3	1.89	2.68	.97	.57
Test Items	5	2.79	3.42	1.15	1.04
Percentile Ranks	3	1.10	1.42	.79	.67
Measures of Central Tendency	3	.58	1.32	.82	.98
Normal Distribution	3	1.37	2.05	.87	.82
Normal Curve	1	1.00	1.00	.00	.00
Correlation Coefficient	1	.58	.68	.49	.46
Correlation/Scatter Diagram	1	.36	.57	.48	.49
Validity	3	1.16	2.10	.87	.55
Reliability/Factors Affecting	3	1.05	2.42	.60	.82
Reliability/Interpretation	3	1.05	1.74	.89	.85
Standard Error of Measurement	1	.58	.74	.49	.44
Types of Tests	3	2.26	2.31	.78	.73
Test Norms/Intelligence Quotient	3	1.79	2.31	.83	.86
Standardized Test Information	3	2.05	2.16	1.00	.74
TOTAL TEST	42	21.05	28.68	4.22	3.65

Table 1.--Learner performance data--Phase 2

Instructional Module	Number Items	Mean		Standard Deviation	
		Pretest	Posttest	Pretest	Posttest
Pretest/Posttest	3	1.45	1.82	.62	.60
Behavioral Objectives	3	2.14	2.43	.99	.62
Test Items	5	2.96	3.11	.94	.86
Percentile Ranks	3	1.28	1.93	.75	.80
Measures of Central Tendency	3	.82	1.43	.71	1.02
Normal Distribution	3	.71	1.96	.84	.94
Normal Curve	1	.86	1.00	.35	.00
Correlation Coefficient	1	.43	.89	.49	.31
Correlation/Scatter Diagram	1	.25	.64	.43	.48
Validity	3	1.36	2.03	.97	.94
Reliability/Factors Affecting	3	1.57	2.28	.62	.80
Reliability/Interpretation	3	1.28	1.86	1.06	.87
Standard Error of Measurement	1	.64	.89	.48	.31
Types of Tests	3	2.32	2.64	.66	.55
Test Norms/Intelligence Quotient	3	1.46	1.75	.62	.63
Standardized Test Information	3	2.00	2.21	.71	.72
TOTAL TEST	42	21.57	28.89	3.23	4.74

Table 2.--Learner performance data--Phase 4

The similarities which existed between the two implementations of the instructional modules included:

1. Course--The evaluation unit of EED 405--Classroom Organization and Pupil Evaluation was used to implement the instructional modules.

2. Instructor--The same graduate assistant instructor was given complete responsibility for the evaluation unit.

3. Population--The learners were all elementary education majors in either their junior or senior year at The Florida State University.

4. Length of Unit--The evaluation unit was allocated a total of nine one-hour class sessions.

The significant differences between the two implementations of the instructional modules are:

1. Test Items--A set of 42 multiple choice test items was used to measure the learners' performance on the 16 instructional modules which specified multiple choice items as the method of evaluation. There were 3 test items replaced and 11 test items modified during the revision of the instructional materials.

2. Testing Procedures--The time between the pre- and posttest was reduced from 16 calendar days during Phase 2 to 8 calendar days during Phase 4.

3. Sample Size--Nineteen learners participated in Phase 2 and 28 learners participated in Phase 4 of the PIERIM model.

Interpretation of Learner Performance

The learners' performance can be expected to deviate from the performance predicted by criterion-referenced measurement and mastery models of learning to the extent that the following assumptions, implicit in the procedures used to design and/or implement the instructional modules and tests, are violated:

1. Learners enter the instructional system in an unlearned state.
2. Learners, who interact with the instructional resources specified, change from an unlearned to a learned state.
3. Learners possess any prerequisite competencies required to interact with the instructional resources that are identified for the instructional modules.
4. Learners have sufficient time to achieve mastery on each instructional module.
5. Test items, for each instructional module, represented a homogeneous sample of the performance described by the behavioral objective.

The learners' performance was measured for the set of 16 instructional modules using the same form of a 42 item multiple choice test as both the pre- and posttest in a One Group Pretest-Posttest Design. Revisions were made to the test during Phase 3 of the PIERIM model and this factor should be considered when comparing the performance of Group 1 (i.e., Conventional Classroom Group) and Group 2 (i.e., Self-Instruction Group). The sample size for Group 1 and Group 2 were 19 and 28 learners respectively.

Violation of Statistical Assumptions

The interpretation of learner performance data is further complicated by the use of intact classroom groups to study the effects of the instructional materials and/or procedures on the learners' performance. The use of intact classroom groups violates one of the basic underlying assumptions of inferential statistics (i.e., random sampling of learners from the population). The assumption that the underlying distribution of the trait being evaluated approximates the normal distribution is violated as the actual effectiveness of the instructional materials and/or procedures approach their theoretical limit of 100 percent effectiveness. Non-parametric statistics were selected for analysis of the learner performance data. Non-parametric statistics (i.e., phi coefficients and McNemar's Test) were selected to be reported by the Instructional Support System (ISS), computer program STAT because there are no assumptions required concerning the underlying distribution of the performance data.

The purpose of designing and implementing the instructional modules in a self-instruction environment was for the learners to achieve at least the standard of performance specified for each of the instructional modules. Learning is inferred from gains in the proportion of learners achieving the standard of performance from pretest to posttest. It is important to remember that the research design utilized (i.e., One Group Pretest-Posttest Design) makes it impossible to separate the gains attributable to the effects of testing from the gains attributable to the instructional treatment. Utilizing the proportion of learners achieving at least the standard of performance on the pretest and posttest the gains from pretest to posttest and the ratio of the gains to potential gain are reported for each instructional module (see Table 3).

Instructional Module	Proportions				
	Pre-Test	Post Test	Gain 1	Gain 2	Ratio Gain 1/Gain 2
Pretest/Posttest	.393	.714*	.321	.607	.528
Behavioral Objectives	.786*	.929*	.143	.214	.668
Test Items	.250	.250	.000	.750	.000
Percentile Ranks	.321	.714*	.393	.679	.579
Measures of Central Tendency	.179	.464	.285	.821	.347
Normal Distribution	.179	.607	.428	.821	.521
Normal Curve	.857*	1.000*	.143	.143	1.000
Correlation Coefficient	.429	.893*	.464	.571	.813
Correlation/Scatter Diagram	.250	.643	.390	.750	.524
Validity	.429	.786*	.357	.571	.625
Reliability/Factors Affecting	.500	.786*	.286	.500	.536
Reliability/Interpretation	.464	.679	.215	.536	.401
Standard Error of Measurement	.643	.893*	.250	.357	.700
Types of Tests	.893*	.964*	.071	.107	.663
Test Norms/Intelligence Quotient	.536	.643	.107	.464	.230
Standardized Test Information	.750*	.821*	.071	.250	.284

Gain 1 - Actual gain in the proportion of learners achieving the standard of performance from pretest to posttest.

Gain 2 - Maximum gain possible in the proportion of learners achieving the standard of performance from pretest to posttest.

* Instructional Module on which at least 70 % of the learners achieved the standard of performance specified for the behavioral objective.

Table 3.--Changes in the proportion of Group 2 learners achieving the standard of performance from pretest to posttest.

Any arbitrary standard can be selected as the performance standard for a system model for design of instruction. For purposes of illustrating the use of a standard of performance for a system model for design of instruction, 70 percent is selected as the system standard for the PIERIM model. The learners achieved the system standard of performance on four of the 16 instructional modules on the pretest and for 10 of the 16 instructional modules on the posttest (see Table 3). There would be reason to suspect that for at least the four instructional modules on which the system standard of 70 percent was achieved on the pretest that the topic had been previously taught in other education courses or the instructional objective was so obvious as not to require instruction. A comparison of the ratios of gains to potential gains requires the assumption that a gain from .80Ø to .90Ø (i.e., .10/.20 = .50Ø) is equivalent to a gain of from .40Ø to .70Ø (i.e., .30/.60 = .50Ø).

Revision Indicator for Instructional Modules

When the instructor of the elementary education course reviewed the set of summary reports produced by the ISS program STAT, he reported that the volume of information contained in the reports was overwhelming. It was determined that a single rank indicator for each instructional module would be an asset to the instructor and educational technologist by directing their efforts during the revision of the instructional modules. Neither the summary reports produced by the computer programs nor the Revision Indicator have actually been utilized to support Phase 3 of the PIERIM model.

The rationale for the Revision Indicator was to select a number of statistics, which were available to the instructor and educational technologist, and predict the direction in which each each statistic would be expected to change on the basis of criterion-referenced measurement and/or mastery models of learning. The Revision Indicator is a single composite value derived from the following statistics:

1. Mean--The posttest mean is predicted to be greater than the pretest mean. The means for Group 1 and Group 2 (see Tables 1 and 2) indicate that the mean of each instructional module did in fact increase from pretest to posttest.

2. Standard Deviation--The posttest standard deviation is predicted to be less than the pretest standard deviation. The standard deviations for Group 1 (see Table 1) and Group 2 (see Table 2) indicate that for some of the instructional modules the standard deviations changed in the opposite direction.

3. Maximum Pretest Score--Learners who achieve a maximum score on the pretest are predicted to achieve mastery on the posttest.

4. Posttest Scores of Zero--Less than 5% of the learners are predicted to be in an unlearned state on each of the items related to an instructional module.

5. Phi Coefficients--Each of the inter item phi coefficients for a set of items related to an instructional module are predicted to be positive. The total number of negative phi coefficients is calculated for the set of items for each instructional module.

6. Proportion of Correct Answers--The proportion of learners who answered an item correctly on the posttest is predicted to be greater than .50.

7. Alternatives for Test Items--It is predicted that on the pretest, at least one learner will select each alternative of the multiple choice items.

8. Posttest Performance--When the group of learners are divided into upper and lower 50%, on the basis of total test score, at least 80% of the learners in the upper 50% are predicted to answer the item correctly.

9. Fail/Fail Category of Performance--The mean proportion of the learners in the fail/fail category of performance was calculated for Group 1 and Group 2 and each was found to approximate .25. The proportion of learners in the fail/fail category is predicted to be less than .25.

Instructional modules and/or test items are categorized as positive (+) if there is agreement between the observed and predicted direction of change. The instructional modules and/or test items are categorized as negative (-) if there is disagreement between the observed and predicted direction of change. The negative indicators are totaled for each instructional module and the total is referred to as the Revision Indicator.

3. RESULTS

Using the performance data for Group 1 and Group 2, Revision Indicators were calculated for each instructional module (see Table 4). There is substantial agreement between the rank-

Instructional Module	GROUP 1	GROUP 2
Pretest/Posttest	12	10
Behavioral Objectives	3	5
Test Items	11	10
Percentile Ranks	10	6
Measures of Central Tendency	13	12
Normal Distribution	5	8
Normal Curve	2	1
Correlation Coefficient	3	2
Correlation/Scatter Diagram	4	5
Validity	5	7
Reliability/Factors Affecting	7	7
Reliability/Interpretation	6	6
Standard Error of Measurement	2	3
Types of Tests	5	3
Test Norms/Intelligence Quotient	7	7
Standardized Test Information	5	7

Numbers represent the total number of negative (-) indicators for an instructional module

Group 1 represents the 19 learners who participated in Phase 2
Group 2 represents the 28 learners who participated in Phase 4

Table 4.--Revision Indicators for instructional modules

ings of the instructional modules using the Revision Indicators derived from the learner performance data for Group 1 and Group 2 ($r_s = .83$). The same three instructional modules and related test items--Measures of Central Tendency, Pretest/Posttest, and Test Items--were identified as being in need for review and possible revision. The Pretest/Posttest instructional module was the only one of the three instructional modules identified which had actually been revised during Phase 3 of the PIERIM model.

4. IMPLICATIONS FOR FUTURE RESEARCH

The preliminary work related to the development of the Revision Indicator provides one method of ranking instructional modules which are evaluated using criterion-referenced measures. The methodology appears to be one method of better utilizing data derived from learner performance during the formative evaluation of instructional materials.

There is a need for the development of a simplified method of ranking instructional modules as to their relative need for revision and a rationale for terminating the revision process for an individual instructional module. The preliminary work related to the Revision Indicator could possibly be expanded to include subjective ratings by the instructor and/or the learners. Research related to the use of minimum change values in the calculation of the Revision Indicator rather than the simpler dichotomy which classifies observed changes as being either in a specified direction or in the opposite direction could possible improve the sensitivity of the Revision Indicator.

A rationale is needed for selecting the criteria to be used to terminate the revision process for an instructional module. Should the criteria be the same for instructional modules produced by a selection model and a design model? The criteria of available time and financial resources between successive implementations of the instructional modules must be considered when the design goals of an instructional system are established.

REFERENCES

Bloom, B. S. Learning for mastery. U.C.L.A. center for the study of evaluation of instructional programs: Evaluation comment, 1968, 1(2), 1-12.

Bloom, B. S. Some theoretical issues relating to educational evaluation. In R. W. Tyler (Ed.) Educational evaluation: New roles, new means. Sixty-eighth yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1969. Pp. 26-50.

Briggs, L. J. A handbook of procedures for the design of instruction. Pittsburgh, Pa.: American Institutes for Research, 1970.

Cox, R. C. & Vargas, J. S. A comparison of item selection techniques for norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, February, 1966.

Cronbach, L. J. Evaluation for course improvement. Teachers College Record, 1963, 64, 672-683.

Deterline, W. A. Practical problems in program production. In P. C. Lange (Ed.), Programmed instruction. Sixty-sixth yearbo k of the National Society for the Study of Education, Part II. Chicago: University of Chicago, 1967. Pp. 178-216.

Dick, W. Some directions for the College of Education in the 1970's. In D. Hansen, W. Dick & H. T. Lippert, Annual progress report, January 1, 1968 through December 31, 1968.

Ebel, R. L. Evaluating content validity. Educational and Psychological Measurement, 1956, 16, 269-282.

Ebel, R. L. Measurement and the teacher. Educational Leadership, 1962, 20, 20-24(43).

Evans, J. Behavioral objectives are no damn good. In Aerospace Education Foundation Techology and innovation in education. New York: Praeger, 1968. Pp. 41-45.

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. American Psycholgist, 1963, 18, 519-521.

Glaser, R. Toward a behavioral science base for instructional design. In R. Glaser (Ed.), Teaching machines and programmed learning, II. Washington: National Education Association, 1965. Pp. 771-809.

Glaser, R. Objectives and evaluation: An individualized system. Science Education News, June 1967, 1-3.

Glaser, R. Evaluation of instruction and changing educational models. CSEIP Occasional Report No. 13. Los Angeles: University of California, 1968.

Glaser, R., & Cox, R. C. Criterion-referenced testing for the measurement of educational outcomes. In R. A. Weisgerber (Ed.), Instructional process and media innovation. Chicago: Rand McNally, 1968. Pp. 545-550.

Hills, J. R. Experience in small graduate classes and approaches to evaluating criterion-related measures. In C. McGuire (Chm.) Criterion related measures: Bane or boon? Symposium presented at the annual meeting of the American Educational Research Association, Minneapolis, March, 1970.

Husek, T. R. Different kinds of evaluation and their implications for test development. UCLA CSE Evaluation Comment, 1969, 2,(1), 8-10.

Lindvall, C. M., & Cox, R. C. The role of evaluation in programs for individualized instruction. In R. W. Tyler (Ed.), Educational evaluation: New roles, new means. Sixty-eighth yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1969. Pp. 156-188.

Lipe, J. G. The development and implementation of a model for the design of individualized instruction at the university level. Unpublished doctoral dissertation, The Florida State University, 1970.

Lumsdaine, A. A. Assessing the effectiveness of instructional programs. In R. Glaser (Ed.), Teaching machines and programmed learning, II. Washington, D. C.: National Education Association, 1965. Pp. 267-320.

Merwin, J. C. Historical review of changing concepts of evaluation. In R. W. Tyler (Ed.), Educational evaluation: New roles, new means. Sixty-eighth yearbook of the National Society for the Study of Education. Chicago: University of Chicago, 1969. Pp. 305-334.

Moxley, R. A. A source of disorder in the schools and a way to reduce it: Two kinds of tests. Educational Technology; Teacher and Technology Supplement, 1970, 1(1), S3-S7.

Popham, W. J. Indices of adequacy for criterion-referenced test items. Paper presented at the meeting of The American Educational Research Association, Minneapolis, March, 1970.

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6(1), 1-9.

Provus, M. Evaluation of ongoing programs in the public schools. In R. W. Tyler (Ed.), Educational evaluation: New roles new means. Sixty-eighth yearbook of the National Society for the Study of Education. Chicago: University of Chicago Press, 1969. Pp. 244-283.

Scriven, M. The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Ed.), Perspectives of curriculum evaluation. Chicago: Rand McNally, 1967. Pp. 39-83.

Stake, R. E. Countenance of educational evaluation. Teachers College Record, 1967, 68, 523-540.

Tyler, R. W. Basic principles of curriculum and instruction. Chicago: University of Chicago Press, 1949.

Tyler, R. W. The function of measurement in improving instruction. In E. F. Lindquist (Ed.), Educational measurement. Washington, D. C.: American Council on Education, 1951. Pp. 47-67.

Wittrock, M. C. The evaluation of instruction. UCLA CSE Evaluation Comment, 1969, 1(4), 1-7.